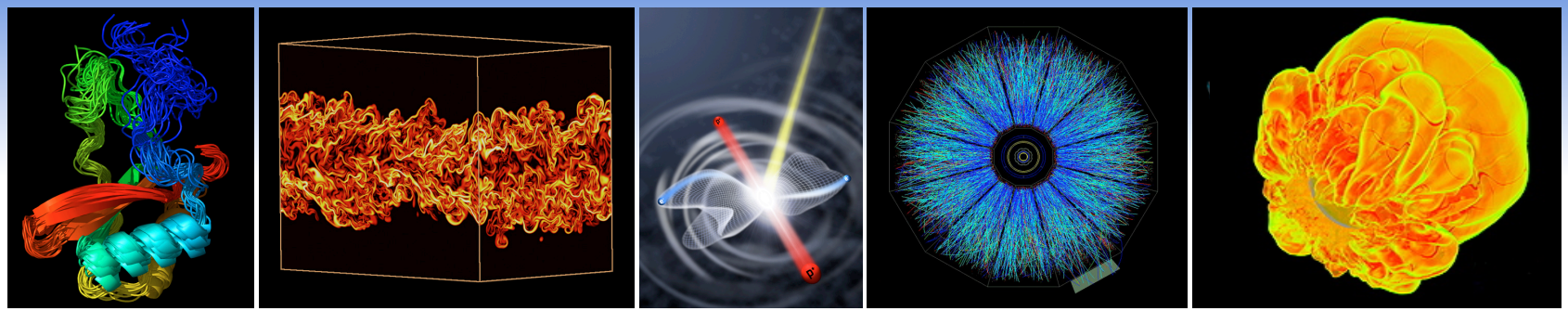


Flash Storage at NERSC

Shane Canon and Jason Hick

Lawrence Berkeley National Laboratory

Lauren Smith Visit
July 2009

BERKELEY LAB

NERSC

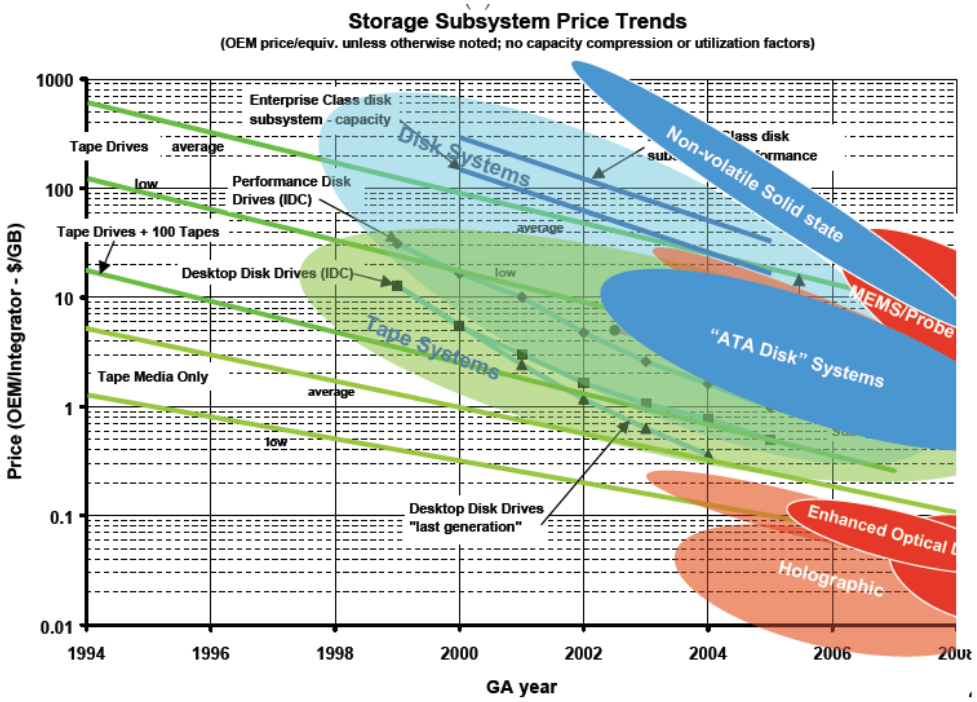
Office of Science
U.S. DEPARTMENT OF ENERGY

Outline

- Technology Introduction (Jason)
- Testing Results (Shane)
- Unfunded Plans (Shane)

Flash Technology Trends

- Solid state storage expected to match disk storage in $/GB by 2014 timeframe
- However, impact could be felt sooner. (Will R&D investment levels in magnetic media continue if SSDs take over consumer market?)
- $/IOPS is competitive, especially for ~1 TB- solutions
- Phase Change Memory could replace NAND (faster and more reliable)
- Probe memory could enter in commodity marketplace (higher bit density, higher latency)



Office of Science
U.S. DEPARTMENT OF ENERGY

*Chart courtesy of Bob Raymond, Sun Microsystems, Pictures courtesy of FusionIO website

Berkeley Lab

Flash Terminology

- NAND - not AND (commodity, error prone)
- NOR - not OR (ROMs, error free)
- SLC - single-level cell, single bit per cell, for write intensive workloads, $30/GB
- MLC - multi-level cell, multiple bits per cell, for read intensive workloads, $5/GB
- Devices manages data in pages and blocks. Pages are typically 512 byte or multiples thereof. Experienced limitations with other than 512 byte pages doing direct I/O with a database. Using with file systems, no problem.

***Chart courtesy of Bob Raymond, Sun Microsystems, Pictures courtesy of FusionIO website**

Not without Challenges

Reliability Overhead

- Where remapping logic is completed (hardware, software)
- Wear issues (10k cycles for MLC, 100k cycles for SLC)

Erasure Overhead

- Writes require erase (slow). Erases must be done in blocks. This leads to trouble with non-sequential I/O as the card fills up. Different cards use different grooming techniques (some CPU intensive - 60% for one card)

NAND SSDs don't act like disks or RAM

- Decades of software development to deal with idiosyncrasies of disk. Similar investments required for solid state storage
- Chip failure requires card replacement (implies mirroring across cards)

Platform Support

- Linux and Windows

Many Advantages

- Minimal latency (50us) for random read relative to disk (10ms)
- Low power (both when active or idle). Several Watts for disk to fraction of Watts for Flash (.15 - .40)
- No mechanical parts to fail. MTBF are comparable to RAM
- Rebuild times on cards should be minutes as opposed to hours with disk
- Peak IOPS (especially with power and form factor), 100K for SLC Flash cards, less than a thousand per disk spindle

Potential Impacts Today

Leverage Flash's IOPS and random read performance to accelerate some workloads

- File System Metadata
- Databases
- Out-of-core
- Data Intensive Applications with heavy random I/O (genomics, Graphs)

Testing to Date

- Metadata backing store for GPFS
- HPSS Metadata backing store
  - HPSS metadata is a large OLTP database (100GB - 1TB)
  - DB2 backups/restore
- Low-level benchmarks
  - iozone
  - xdd

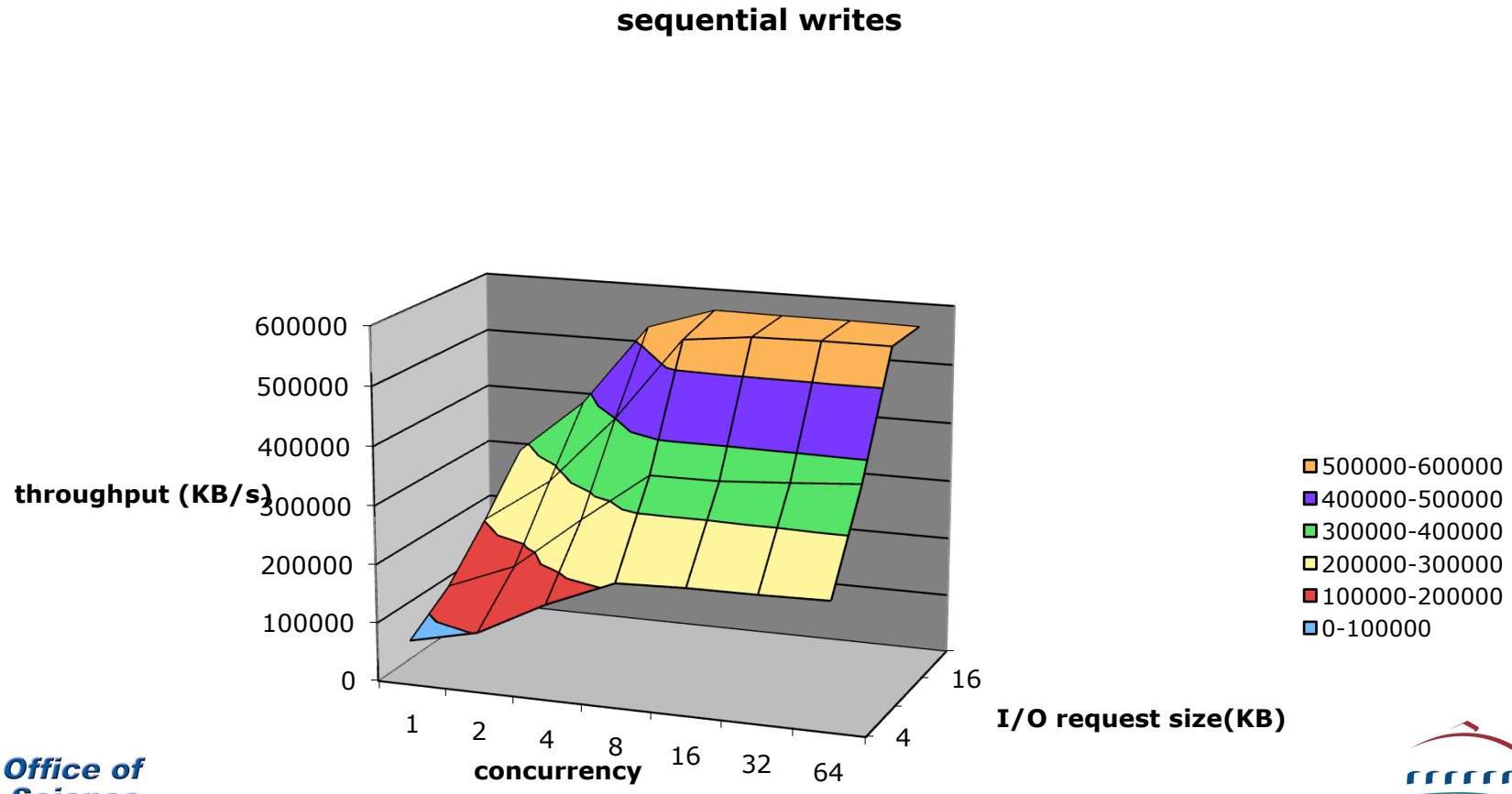
Current Test Resources

Flash Storage Cards

- Loaner Texas Memory Systems card 450 GB SLC (Eval card - $15k)
- 2 FusionIO 160 GB SLC cards ($7.2k/ea) - Higher capacity cards are available now

Current Benchmarks (TMS Ramsan 20 450GB)

For sequential writes, 8KB+ I/O size and 4 way concurrency to get max BW of 600MB/s

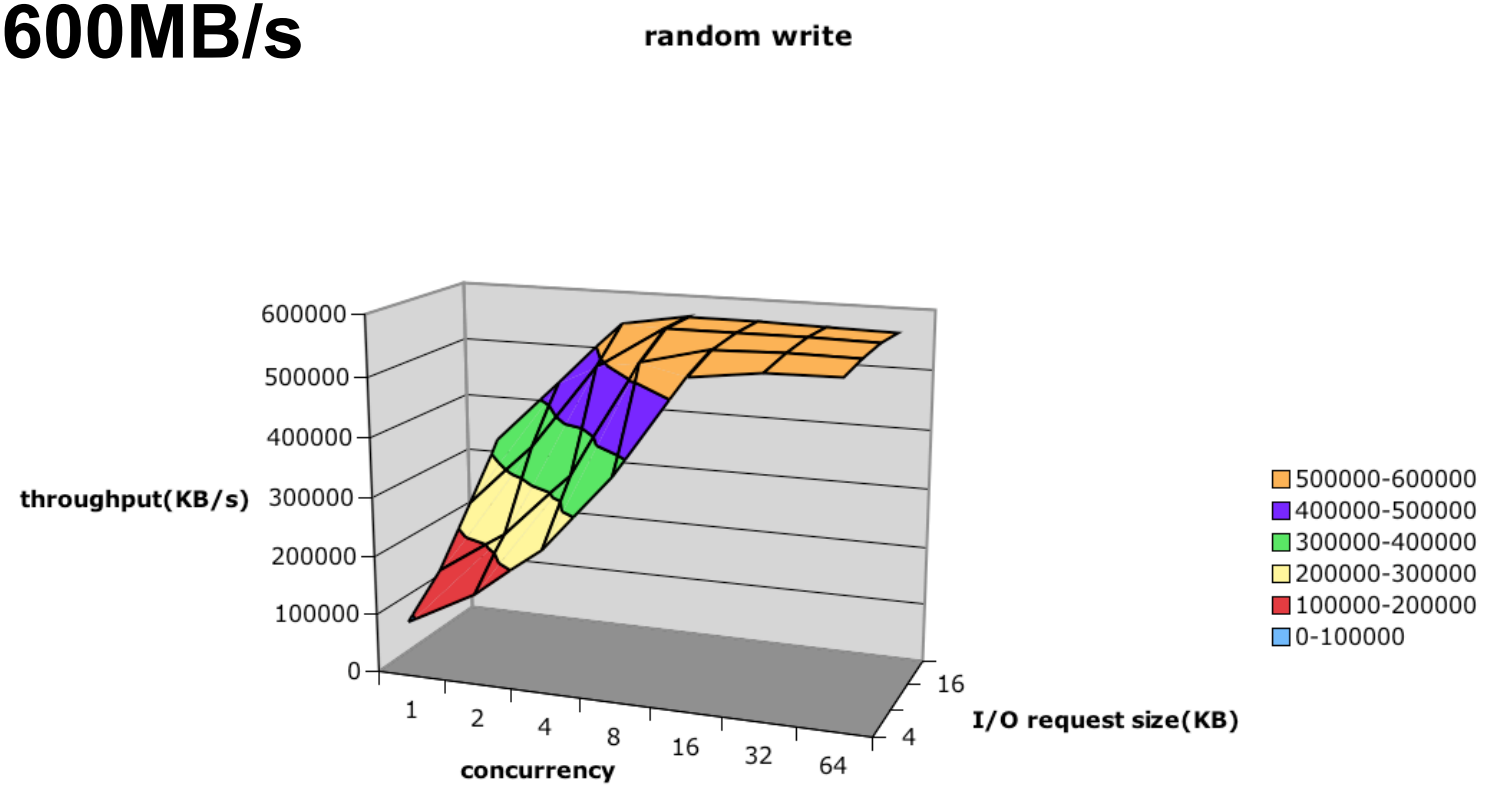


Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

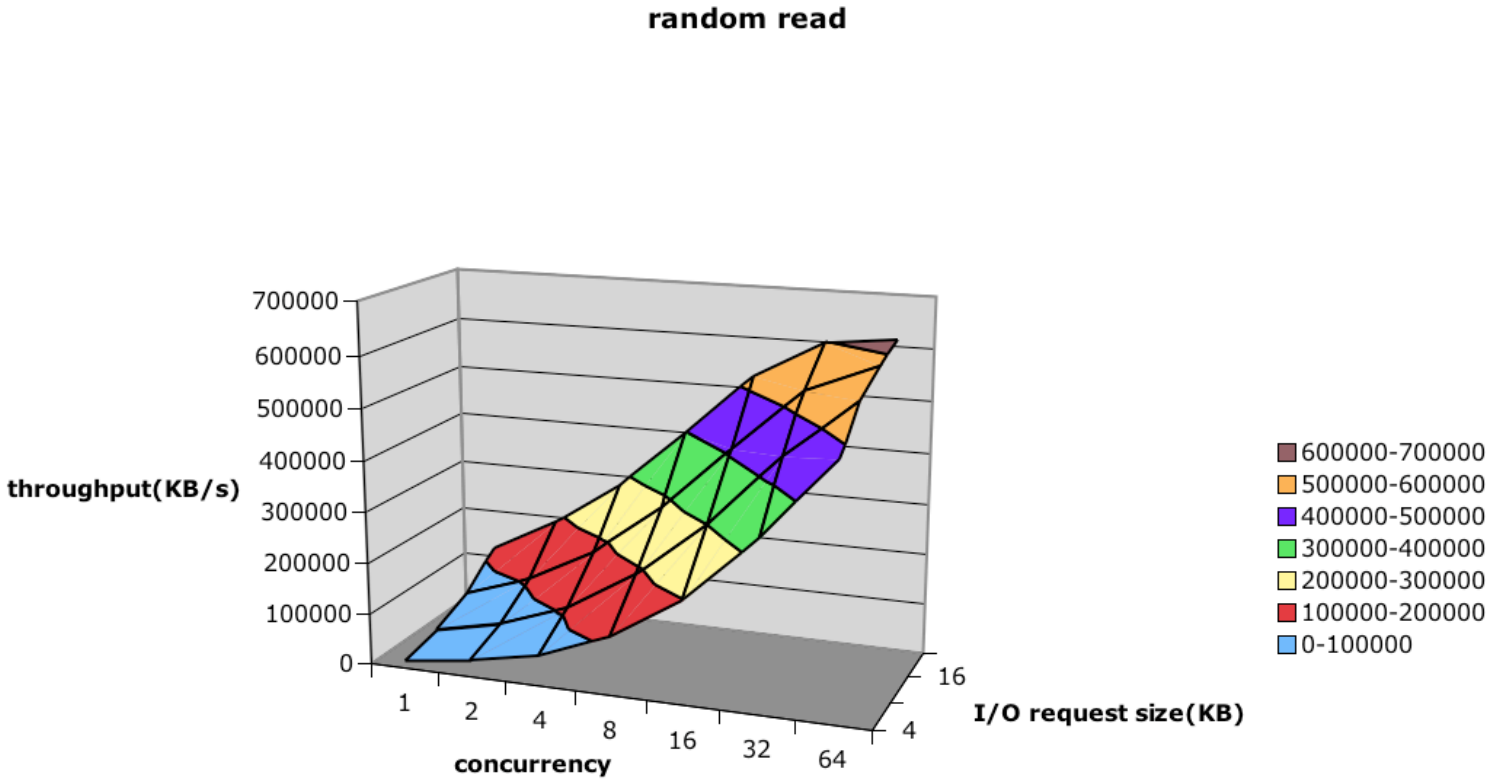
Current Benchmarks (TMS Ramsan 20 450GB)

For random writes and sequential rewrites, 4KB+ I/O size and 4+ way concurrency to get max BW of 600MB/s



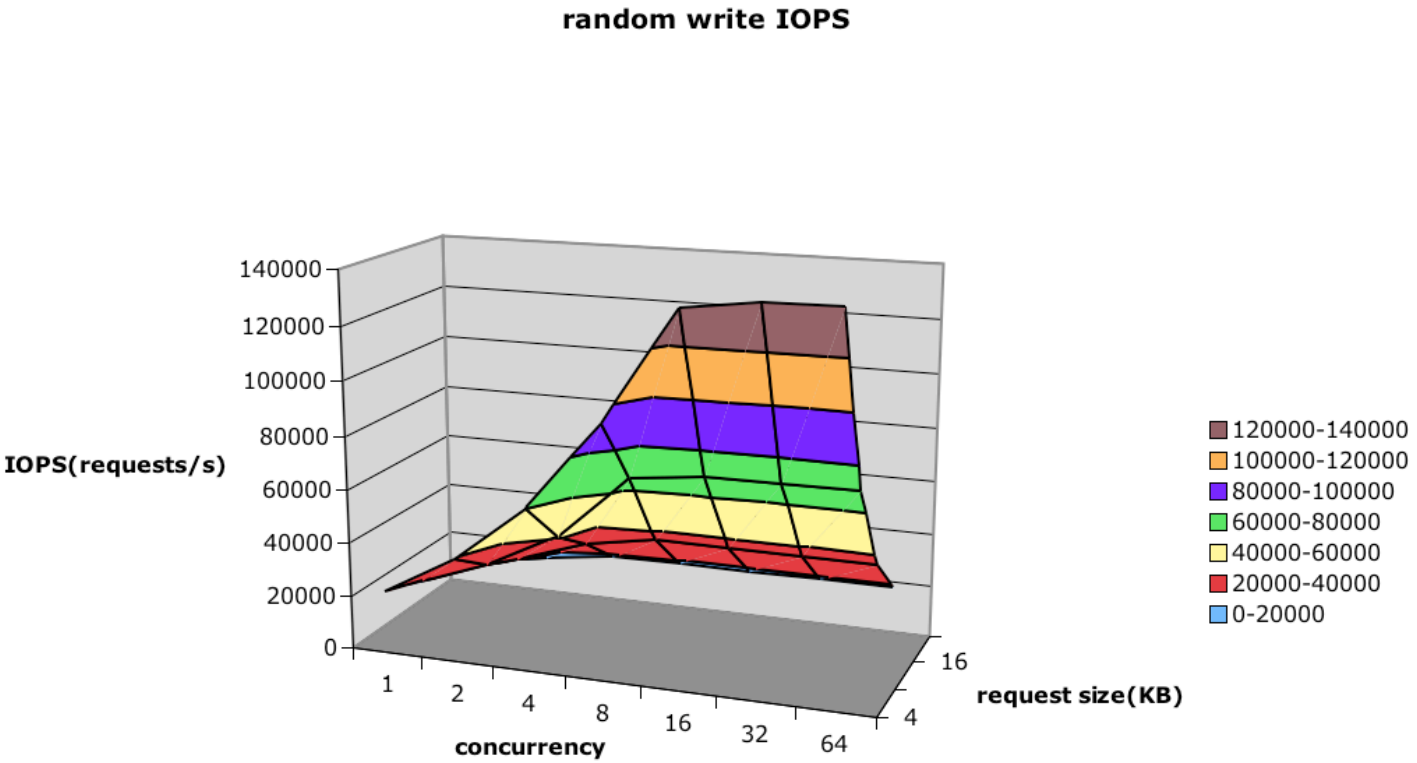
Current Benchmarks (TMS Ramsan 20 450GB)

For random reads, max BW 600 MB/s requires high concurrency of 64 readers



Current Benchmarks (TMS Ramsan 20 450GB)

For random write IOPS, 120,000 IOPS achievable with 8+ writers doing 4K writes

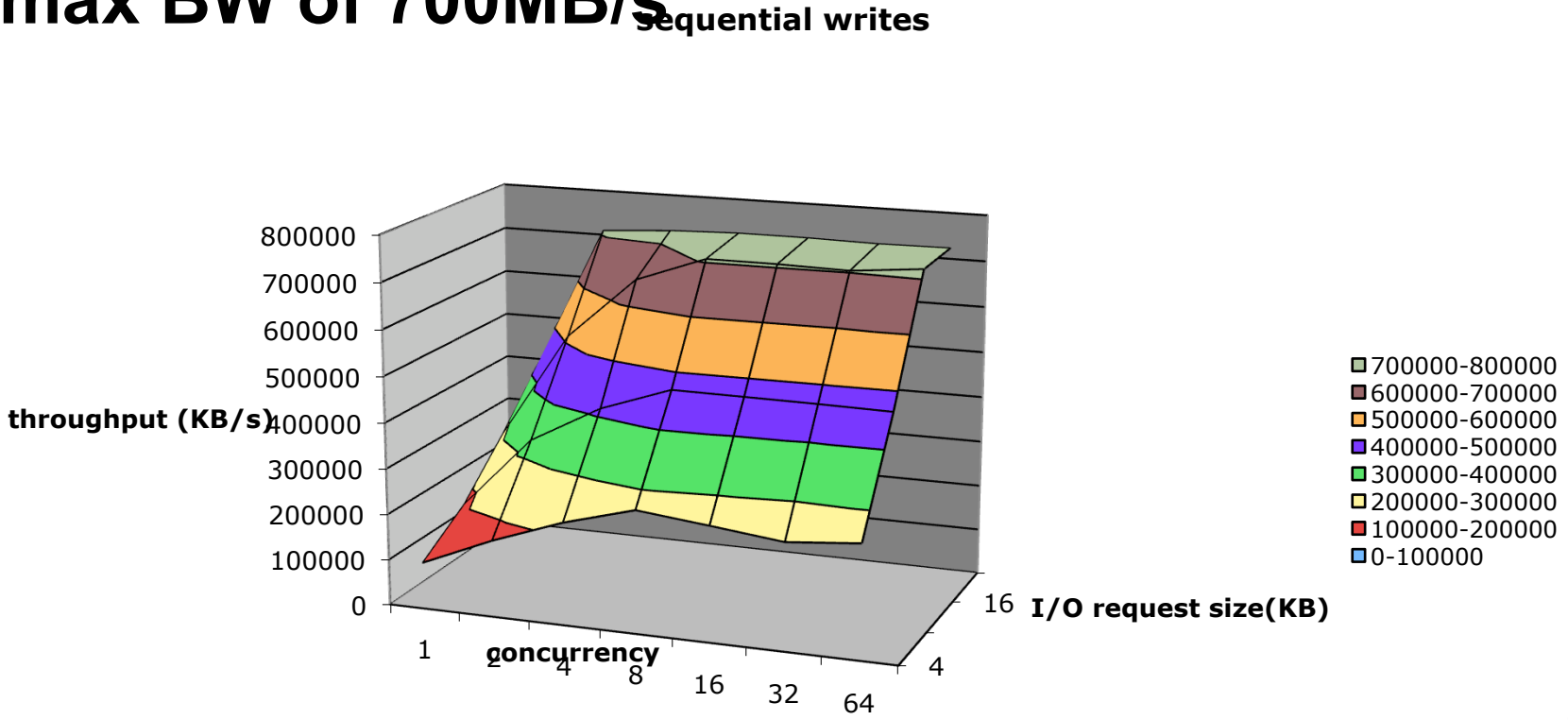


Current Benchmarks (TMS Ramsan 20 450GB)

For random read IOPS, 120,000 IOPS achievable only with 64 readers doing 4K reads

Current Benchmarks (FusionIO 160GB)

For sequential and random writes and rewrites, 16KB+ I/O size and 2 way concurrency to get max BW of 700MB/s



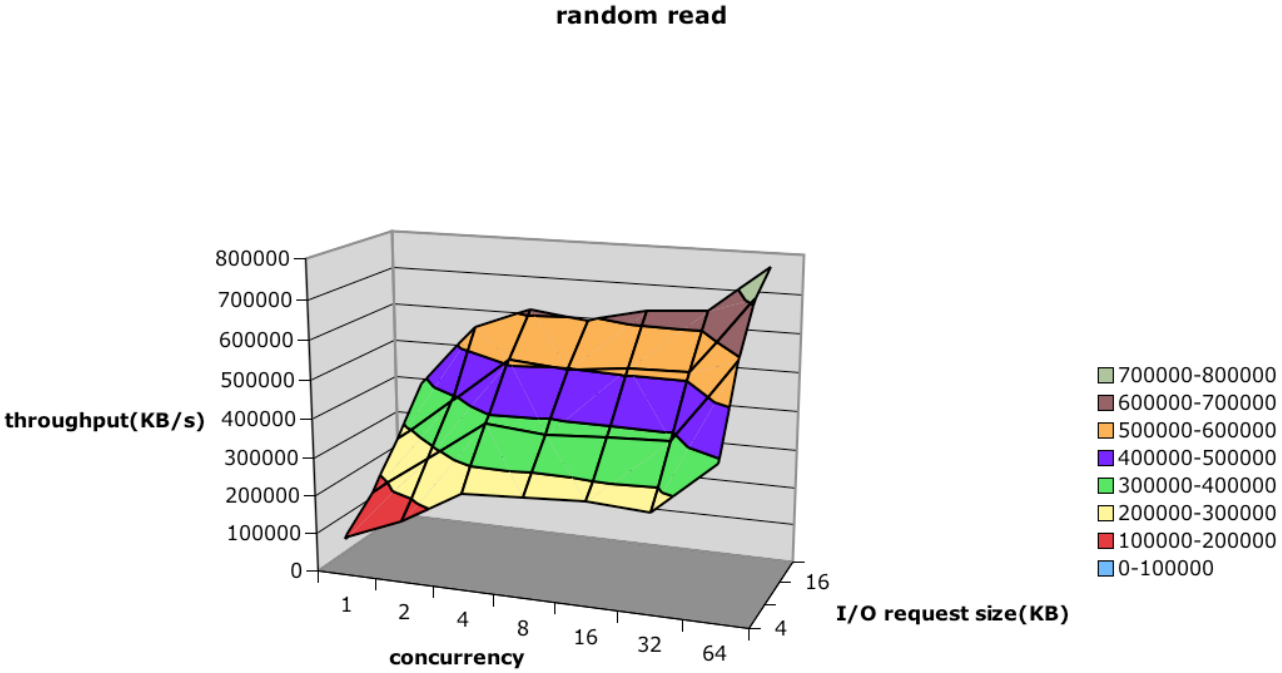
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

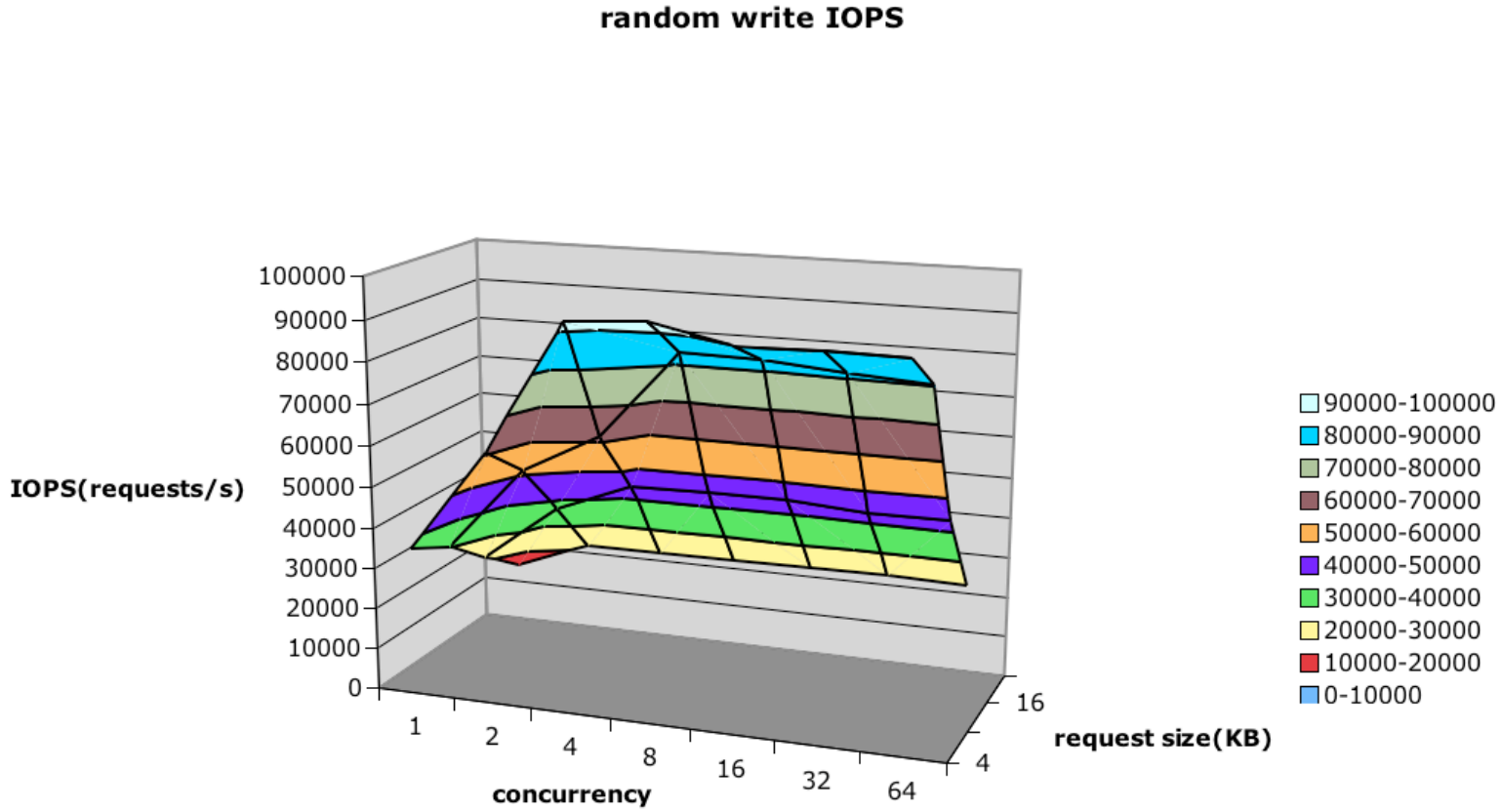
Current Benchmarks (FusionIO 160GB)

For random reads, 16KB+ I/O size and 64 way concurrency to get max BW of 700MB/s



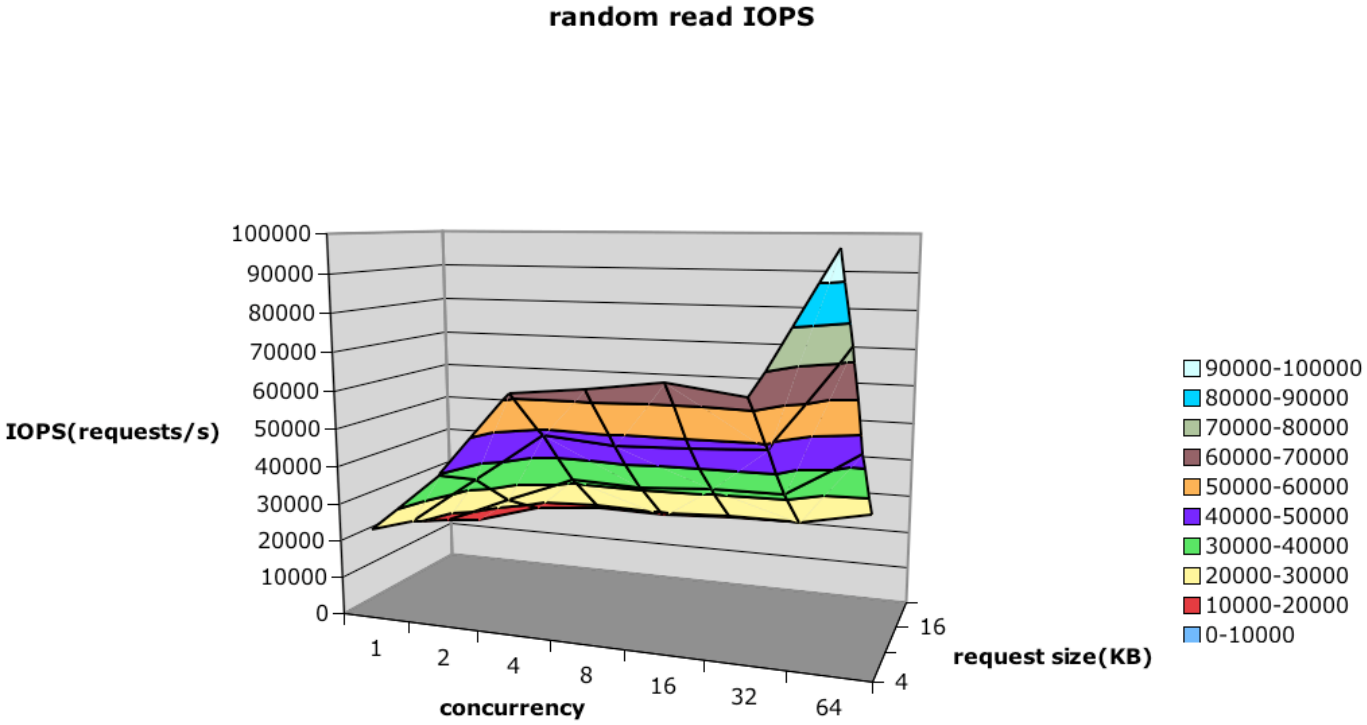
Current Benchmarks (FusionIO 160GB)

For random write IOPS, 4KB I/O size and 4 way concurrency to get max IOPS of 80K



Current Benchmarks (FusionIO 160GB)

For random read IOPS, 4KB I/O size and 64 way concurrency to get max IOPS of 100K



Other Studies

- GPFS Metadata backend – Focused on specific issues. Will repeat for a more general workload evaluation
- DB2 database for HPSS Metadata - Underway

Future Ideas

Larger Testbed

- 10 cards (10 GB/s, 5 TB) – $150k
- 10 high-performance nodes with QDR IB - $150k

Future Research Topics

- Flash based disk pool for GPFS
- Analytics workloads
  - Visualization
  - Data Mining
  - Integration with Hadoop
  - Out-of-core applications/swap
- Databases (Online Transaction Processing)
- New File System Approaches
  - Log Structured FS (NILFS, PLFS)
  - Additional storage hierarchy

Flash Storage in an Exascale Architecture

- Flash landed on Motherboard (low power, inexpensive) – Accelerate Checkpoints, Extend main memory, replace local disk
- Flash in Storage Arrays at the interconnect edge – First level cache to deal with extreme I/O burst. Stream to flash then reorder for sequential friendly storage (i.e. disks). Lower power than 100,000 (or more) spindles
- Flash in Metadata storage